

# When Do Phylogenetic Mixture Models Mimic Other Phylogenetic Models?

Elizabeth S. Allman<sup>1</sup>, John A. Rhodes<sup>1\*</sup>, Seth Sullivan<sup>2</sup>

<sup>1</sup>*Department of Mathematics and Statistics, University of Alaska Fairbanks, Box 756660, Fairbanks, AK, 99775*

<sup>2</sup>*Department of Mathematics, North Carolina State University, Box 8205 Raleigh, NC 27695*

*\*To whom correspondence should be addressed;*

*E-mail: j.rhodes@alaska.edu*

**Abstract**— Phylogenetic mixture models, in which the sites in sequences undergo different substitution processes along the same or different trees, allow the description of heterogeneous evolutionary processes. As data sets consisting of longer sequences become available, it is important to understand such models, for both theoretical insights and use in statistical analyses. Some recent articles have highlighted disturbing “mimicking” behavior in which a distribution from a mixture model is identical to one arising on a different tree or trees. Other works have indicated such problems are unlikely to occur in practice, as they require very special parameter choices.

After surveying some of these works on mixture models, we give several new results. In general, if the number of mixture components is not too large, and we disallow zero or infinite branch lengths, then mimicking does not occur. On the other hand, if the mixture model is locally over-parameterized, it is possible for a phylogenetic mixture model to mimic distributions of another tree model. Though theoretical questions remain, these sorts of results can serve as a guide to when the use of mixture models in either ML or Bayesian frameworks is likely to lead to statistically consistent inference.

Keywords: Phylogenetic mixture models, parameter identifiability, heterogeneous sequence evolution

Model-based phylogenetic inference from sequence data requires compromises between simplicity and biological realism. Typical current modeling assumptions include that all sites evolve on a single tree, according to the same substitution process, often with a simple  $\Gamma$ -distributed scaling of rates across the sites. While one can easily formulate models allowing more complexity, the additional parameters this introduces can be problematic. Not only is software likely to require longer run-times, but one also risks ‘overfitting’ of finite data sets and thus interpreting stochastic variation as meaningful signal. In more extreme cases, over-parameterization can lead to loss of identifiability of some parameter of interest, and thus the loss of statistical consistency of inference as well.

Nonetheless, it is not hard to conceive of data sets for which the modeling assumptions underlying a routine analysis are strongly violated. For instance, different parts of a single gene sequence might undergo rather different substitution processes, perhaps due to different substructures of the protein they encode. Alternatively, lateral transfer of genetic material may have resulted in sequences that are amalgams of those evolving on different trees. Analyzing such data under a standard model simply assumes that neither of these has occurred, and so uses a *misspecified* model. While one would hope there would be some indication of this as the analysis is conducted — perhaps by a poor likelihood score or poor convergence of a Bayesian MCMC run — there is no guarantee of an obvious sign of a problem.

An alternative is to consider *mixture models*, which explicitly allow for such heterogeneity in the data. Mixtures consider several classes of sites which might each evolve according to a distinct process, either on the same tree (a *single-tree mixture model*), or on possibly different trees (a *multitree mixture model*). In both cases the use of a mixture model differs from a partitioned analysis of data, in which the researcher imposes a partitioning of the sites into classes, each of which must evolve according to a single standard model. For a mixture

model, there is no *a priori* partitioning; the proportion of sites in each class is a parameter of the model, and is thus to be inferred.

The single-tree GTR+ $\Gamma$ (+I) model is a familiar, but highly restricted type of mixture, with few parameters, that is commonly used in data analysis. Only recently Chai and Housworth (2011) completed a rigorous proof that the parameters of this model, including the tree topology, are identifiable from its probability distributions in most cases, and thus that it gives consistent inference under maximum likelihood. However, the special case of the F81+ $\Gamma$ +I submodel remains open (Steel, 2009).

On the other hand, a single-tree rate-variation model in which the rate distribution was allowed to be arbitrary was one of the earliest mixture models seen to be problematic, as every tree can produce the same distribution of site patterns (Steel et al., 1994). The no-common-mechanism (NCM) model introduced by Tuffley and Steel (1997) provides another example of a mixture in which distributions do not identify trees. However, these models are rather unusual, in that the number of their parameters grows with sequence length. This extreme over-parameterization is well understood, as is the implication that these models do not lead to statistically consistent inference under a maximum likelihood framework. (Steel (2011) offers a more complete and subtle discussion of NCM models and inference.) Of course these models were introduced to elucidate theoretical points, and were not intended for data analysis.

Much recent work on mixture models has focused on those with a finite (though perhaps large) number of mixture components, allowing more heterogeneity among the classes than the simple scaling of the rate variation models. Several papers have shown that inference from data generated by a mixture process can be poor if the analysis is based on a misspecified non-mixture model (Kolaczowski and Thornton, 2004; Mossel and Vigoda, 2005, 2006). The examples in these works indicate that we may be misled if we ignore the possibility of such heterogeneity. This point is further underscored by Matsen and Steel (2007), who discuss why analysis with a misspecified non-mixture can lead to erroneous inference in some

specific circumstances. As there is no general reason why one should expect good inference with a misspecified model, to our mind these works primarily indicate the importance of further study of mixture models, so that they may be applied intelligently when substantial heterogeneity is possibly present.

However, several works have indicted that models with a finite number of mixture classes may have theoretical shortcomings as well. Working with no restriction on the number of classes, Štefankovič and Vigoda (2007a,b) emphasize that unless a model is special enough that there are linear inequalities (which they call *linear tests*) distinguishing between unmixed distributions arising on different trees, then there will be cases in which trees can not be identified from single-tree mixture distributions. Matsen et al. (2008) explore this more particularly for the Cavender-Farris-Neyman (CFN) 2-state symmetric model. Considering models with a small number of mixture classes, Štefankovič and Vigoda (2007a,b) and Matsen and Steel (2007) give explicit examples of parameter choices in certain 2-class CFN single-tree mixture models that lead to exactly the same unmixed probability distributions as standard models on a different tree. Such non-identifiability of the tree, or ‘mimicking’ as it is termed by Matsen and Steel (2007), means that the true tree parameter cannot be determined even from an exact theoretical probability distribution, much less from data sampled from it. If the problems highlighted in these works were widespread, then we would be severely limited in our ability to detect heterogeneous process. Moreover, heterogeneous processes on one tree might routinely mislead us into thinking data arose on a different tree. We have encountered researchers who, not surprisingly, found these possibilities quite alarming.

While there is no doubt that certain mixtures are problematic, whether this is really of great practical concern is in fact not at all clear from the results mentioned so far. Thoughtful use of mixture models for data analysis has seemed to perform well for a number of research groups (Ronquist and Huelsenbeck, 2003; Pagel and Meade, 2004, 2005; Le et al., 2008; Wang et al., 2008; Evans and Sullivan, 2012). While publication bias against failed

analyses could be responsible for a lack of reports of difficulties with mixture models in the literature, we also have not heard of such problems through our professional interactions.

Several papers (Allman and Rhodes, 2006; Allman et al., 2011; Rhodes and Sullivant, 2012) have given a strong theoretical indication that problematic mixtures, for which trees are non-identifiable, are quite rare. Using algebraic techniques building on the idea of phylogenetic invariants, these works show in a variety of contexts that mixture distributions cannot mimic distributions arising on other trees, for generic choices of numerical parameters. ‘Generic’ here has a precise meaning that informally can be expressed as “if the model parameters are chosen at random, and thus do not have any special values or relationships among themselves.” More formally, the set of exceptional parameters leading to non-identifiability is of strictly smaller dimension than the full parameter space. Thus if the true parameters were chosen by throwing a dart at the parameter space, they would be sure to lie off that exceptional set. Rhodes and Sullivant (2012) give an upper bound on the number of classes that, for a quite general model, ensures generic identifiability of the trees in all single-tree and in many multitree mixtures. This bound is exponential in the number of taxa, and likely to be larger than the number of classes one would actually use in data analysis.

While these positive theoretical results indicate one should seldom encounter problems with the judicious use of a mixture model in data analysis, one may still worry about the possible exceptions. The exceptional cases are generally not explicitly characterized in these papers, and the arguments used to establish that they form a set of lower dimension are rather technical. The intuition of the authors is that the potential exceptional set one could extract from these works are likely to be much larger than the true exceptional set, as an artifact of the techniques of proof. Moreover, experience with other types of statistical models outside of phylogenetics (*e.g.*, hidden Markov models, Bayesian networks) with similar exceptional sets of non-identifiability has shown they can still be quite useful, and are generally not problematic for data analysis.

In this work we first give mathematical justification — with no cryptic assumptions of genericity of parameters — that a limited amount of heterogeneity in a single-tree mixture cannot mimic evolution on a different tree in most relevant circumstances. We also show how examples of non-identifiability of trees due to mixture processes can arise from a readily understood issue of *local over-parameterization*. This explains the 2-class mimicking examples of Štefankovič and Vigoda (2007a,b) and Matsen and Steel (2007), which are constructed for models whose parameter space is of larger dimension than the distribution space for a 4-taxon tree. However, this is not the setting in which most data analysis is likely to take place. For 4-state models encompassing those such as the general time-reversible (GTR) which are in common use, we show even 3-component mixtures cannot mimic non-mixtures. While our positive identifiability results do not allow as many mixture components as the ones obtained for generic parameters by Rhodes and Sullivant (2012), by excluding the possibility of exceptions they are, in some sense, more complete. Finally, for certain group-based models (Jukes-Cantor and Kimura 2 parameter), for which linear tests exist, we also obtain results indicating that if mimicking does occur for multitree mixtures, then it is not entirely misleading. In the case of fully-resolved trees, any mimicking distribution can only agree with a distribution coming from one of the topological trees appearing in the mixture.

The description of these results here provides only an outline, as precise statements require a thorough specification of the basic model. We therefore formulate the *general continuous-time model*, and mixtures arising from it, in the next section. Then in a subsequent section a number of theorems formally encapsulate the claims above.

Our results concern only the theoretical limits of what one might be able to infer from a data set; we do not study any issues relating to the performance of mixture models for inference from finite data sets. Our hope is to further dispel concerns that the use of phylogenetic mixture models is inherently problematic, or that inhomogeneous processes are likely to lead to inherently undetectable mistakes in most settings in which inference is performed. As is the case for any statistical model, phylogenetic mixture models must be

applied thoughtfully, and whether they are useful or not will depend on the data.

The mathematical tools we use to obtain our results involve the polynomial equalities called phylogenetic invariants, which have been extensively studied for both the group-based models and the general Markov model, and mixtures built from them. However, we supplement these with some polynomial *inequalities*. While the potential usefulness of inequalities was made clear even in the seminal paper of Cavender and Felsenstein (1987) which introduced invariants, their study unfortunately remains much less developed than the study of invariants. Though a deeper understanding of inequalities for both unmixed and mixture models would be highly desirable, here we make do with a few *ad hoc* ones.

## PHYLOGENETIC MIXTURE MODELS

In this section, we describe the class of phylogenetic models that we study. Our definition of an unmixed phylogenetic model is broad, encompassing most standard phylogenetic models, including those studied by Štefankovič and Vigoda (2007a,b), Matsen and Steel (2007), and Matsen et al. (2008). Informally, we consider continuous-time models, but do not require time-reversibility or stationarity, and allow the substitution process to change at a finite set of points on the tree. Such relaxations of the usual modeling assumptions have appeared in several works (Yang and Roberts, 1995; Galtier and Gouy, 1998; Yap and Speed, 2005).

We assume that the random variables modeling characters have  $\kappa \geq 2$  states, the most important values being  $\kappa = 4$  (DNA models),  $\kappa = 2$  (purine/pyrimidine models), and  $\kappa = 20$  (protein models).

By a *rate matrix* for a state substitution process we mean a  $\kappa \times \kappa$  matrix with nonnegative off-diagonal entries, whose row sums are all zero. (To fix a scaling, one may also impose some normalization convention.) Such a rate matrix  $Q = (q_{ij})$  generates a continuous-time  $\kappa$ -state Markov chain. Associated with  $Q$  is a directed graph,  $G_Q$ , on nodes  $\{1, 2, \dots, \kappa\}$  which has an edge  $i \rightarrow j$  if and only if  $q_{ij} \neq 0$ . The process defined by  $Q$  is *irreducible* if,

and only if,  $G_Q$  is strongly connected, that is, there is a directed path from node  $i$  to node  $j$  for all  $i, j$ . Irreducibility guarantees that for all  $t > 0$  the discrete-time Markov transition matrix  $\exp(Qt)$  has strictly positive entries. Of course  $\exp(Qt)$  is the identity matrix when  $t = 0$ , and so has zero entries.

Consider an unrooted, combinatorial, phylogenetic tree,  $T$ , in which we allow polytomies. Then by the *general continuous-time model* on  $T$ , we mean the following: First, possibly introduce a finite number of degree 2 nodes (in order to model a root, and points where the state substitution process changes) along any of the edges of  $T$  to obtain  $T'$ . Then choose some node to serve as a root of  $T'$ , and make any assignment of a strictly positive  $\kappa$ -state distribution  $\boldsymbol{\pi}$  at the root. Irreducible rate matrices  $Q_i$  and edge lengths  $t_i \in \mathbb{R}_{\geq 0}$  are assigned to each edge  $i$  of  $T'$ . This notion is more general than is often used in most practical data analysis, since 1)  $\boldsymbol{\pi}$  need not be the stationary distribution of any  $Q_i$ , and 2) the  $Q_i$  may be different for each edge; we do not assume a common process across the tree. We at times restrict to considering only irreducible rate matrices of a certain form (*e.g.*, Jukes-Cantor, or GTR) and specialized  $\boldsymbol{\pi}$ , in order to draw conclusions about submodels.

If numerical model parameters are specified as above, then the Markov transition matrix on edge  $i$  of  $T'$  is  $M_i = \exp(Q_i t_i)$ . If  $T''$  denotes the tree obtained from  $T'$  by suppressing non-root nodes of degree 2, and edges  $i, i+1, \dots, i+r$  of  $T'$  become a single edge of  $T''$ , then one defines a Markov matrix on that edge of  $T''$  as the product  $M_i M_{i+1} \cdots M_{i+r}$ . Then one can compute the probability distribution arising from the numerical parameters on  $T'$ , by using the root distribution and edge transition matrices on  $T''$  in the usual way. From the assumption of irreducibility of rate matrices we immediately obtain the following.

**Lemma 1.** *Consider any choice of general continuous-time parameters on a phylogenetic tree  $T$ . Then the Markov transition matrices associated to the edges of  $T'$  and  $T''$  are each either the identity matrix, or a nonsingular matrix with strictly positive entries.*

By  $\mathcal{M}_T$  we denote the set of all probability distributions arising on  $T$  for all choices of general continuous-time parameters. Later in this paper, we use the same notation for



a submodel obtained by restricting parameters to a specific form. If no such restriction is made, we refer to  $\mathcal{M}_T$  as *the general continuous-time model* on  $T$ .

Let  $\mathcal{M}_T^+ \subseteq \mathcal{M}_T$  denote the subset of distributions obtained by requiring that no internal branch lengths are zero, that is all  $t_i \in \mathbb{R}_{>0}$  except possibly for pendent edges. This is the *open phylogenetic model*. Since we allow trees to have polytomies, any distribution in  $\mathcal{M}_T$  is contained in the open model for a possibly different tree; one merely contracts all internal edges of  $T$  which were assigned branch length zero, thus introducing new polytomies.

If  $\mathcal{T} = \{T_1, \dots, T_r\}$  is a multiset of trees, then the *mixture model*

$$\mathcal{M}_{\mathcal{T}} = \mathcal{M}_{T_1} * \dots * \mathcal{M}_{T_r}$$

is the set of probability distributions of the form

$$s_1 p_1 + s_2 p_2 + \dots + s_r p_r,$$

where  $p_i \in \mathcal{M}_{T_i}$  is a probability distribution arising on  $T_i$ , and the  $s_i \geq 0$  are weighting parameters with  $s_1 + s_2 + \dots + s_r = 1$ . The *open mixture model*

$$\mathcal{M}_{\mathcal{T}}^+ = \mathcal{M}_{T_1}^+ * \dots * \mathcal{M}_{T_r}^+.$$

is defined similarly, with  $p_i \in \mathcal{M}_{T_i}^+$ . Note that in the open mixture model we allow mixing parameters to be in the closed probability simplex  $\Delta_r = \{\mathbf{s} \mid s_i \geq 0, \sum s_i = 1\}$ . If all mixing parameters are required to be strictly positive, we denote the set of distributions by  $\mathcal{M}_{\mathcal{T}}^{++}$ , since we are then working with the open probability simplex  $\Delta_r^+ = \{\mathbf{s} \mid s_i > 0, \sum s_i = 1\}$ .

## RESULTS

### *Single-tree Mixture Models*

Matsen and Steel (2007) and Štefankovič and Vigoda (2007b) showed that under the CFN model it is possible for a 2-class mixture on a single tree (that is,  $\mathcal{T} = \{T, T\}$ ) to produce

distributions matching those of an unmixed model on a different tree. Matsen et al. (2008) showed that this is possible if, and only if, the trees involved differ by application of a single NNI move.

Our main result in this setting shows that these possibilities are essentially a “fluke of low dimensions.” In a subsequent section a further analysis will show that this mimicking is a consequence of local over-parametrization, arising essentially because the CFN model is a 2-state model.

**Theorem 2.** *Consider the  $\kappa$ -state general continuous-time phylogenetic model. Let  $\mathcal{T} = \{T_1, T_1, \dots, T_1\}$  be the multiset consisting of  $\kappa - 1$  copies of tree  $T_1$ , and  $\mathcal{S} = \{T_2\}$ . Then  $\mathcal{M}_{\mathcal{T}} \cap \mathcal{M}_{\mathcal{S}}^+ = \emptyset$  unless  $T_1$  is a refinement of  $T_2$ .*

This result was already known to hold for generic choices of parameters in a slightly more general setting (Allman and Rhodes, 2006), so the contribution here is to remove the generic assumption. Note that for the important case of  $\kappa = 4$ , corresponding to DNA models, this implies that we cannot have a two or three class mixture mimic the distribution on a single tree unless we allow zero length branches in the mixture components. This indicates the examples of Matsen and Steel (2007) and Štefankovič and Vigoda (2007a,b) cannot be generalized to 4-state models, without passing to at least a 4-class mixture.

### *Local Over-parametrization*

Note that the examples of Matsen and Steel (2007) and Štefankovič and Vigoda (2007b) are allowed by Theorem 2, since they are constructed for a model with  $\kappa = 2$  and  $\mathcal{T}$  a 2-element multiset. To see why the existence of such examples should not be too surprising, it is helpful to first consider a 4-leaf tree  $T$  and perform a parameter count for the CFN model. A 2-class single-tree mixture on  $T$  can be specified by 11 numerical parameters: for each class there are 5 Markov transition matrices with 1 free parameter each, and 1 additional mixing parameter. However any 4-taxon CFN mixture distribution on any 4-taxon tree lies

in a certain 7-dimensional space, due to the symmetry of the model. Although this does not prove every distribution with such symmetry must arise from this 2-class mixture, the excess of parameters suggests that it is likely that many do. As a result, one suspects at least some non-mixture distributions on different trees are likely to be mimicked by this 2-class mixture. This suspicion is then confirmed by explicit examples.

When a tree has many more leaves, however, a similar parameter count for the 2-class CFN mixture can fail to indicate potential problems, since the number of parameters grows linearly with the number of leaves, while the dimension of the distribution space grows exponentially. However, we show below how one can extend mimicking examples on small trees to larger trees, thus creating what might at first appear to be more unexpected instances of mimicking. We refer to such examples, where mimicking is produced first on a small tree by allowing an excessive number of mixture components, and then extended to larger trees, as arising from *local over-parameterization*. This notion can be used to produce many new examples of the mimicking phenomenon, on single- or multitree mixtures.

We distinguish here between three types of mimicking, of different degrees of severity. For notational convenience we use  $\mathcal{M}_{\mathcal{T}}^*$  to denote any of the models  $\mathcal{M}_{\mathcal{T}}$ ,  $\mathcal{M}_{\mathcal{T}}^+$ , or  $\mathcal{M}_{\mathcal{T}}^{++}$ .

**Definition 3.** A mixture model  $\mathcal{M}_{\mathcal{T}}^*$  *weakly mimics* distributions in  $\mathcal{M}_{\mathcal{S}}^*$  if  $\mathcal{M}_{\mathcal{T}}^* \cap \mathcal{M}_{\mathcal{S}}^* \neq \emptyset$ . A mixture model  $\mathcal{M}_{\mathcal{T}}^*$  *strongly mimics* distributions in  $\mathcal{M}_{\mathcal{S}}^*$  if  $\dim \mathcal{M}_{\mathcal{T}}^* \cap \mathcal{M}_{\mathcal{S}}^* = \dim \mathcal{M}_{\mathcal{S}}^*$ . A mixture model  $\mathcal{M}_{\mathcal{T}}^*$  *completely mimics* distributions in  $\mathcal{M}_{\mathcal{S}}^*$  if  $\mathcal{M}_{\mathcal{S}}^* \subseteq \mathcal{M}_{\mathcal{T}}^*$ .

Thus weak mimicking requires only a single instance of probability distributions arising on  $\mathcal{S}$  and  $\mathcal{T}$  matching, strong mimicking requires a neighborhood of distributions arising on  $\mathcal{S}$  to be matched by ones arising on  $\mathcal{T}$ , and complete mimicking requires every distribution arising on  $\mathcal{S}$  to be matched by one arising on  $\mathcal{T}$ .

Let  $T$  be a tree on the leaf set  $X$ . For each  $x \in X$ , let  $A_x$  be a new, non-empty set of leaves, and let  $B_x$  be a tree with root  $x$  and leaf set  $A_x$ . A set of such trees  $\mathcal{B} = \{B_x : x \in X\}$  is called a set of *fusion ends* for  $X$ . The *fusion tree*  $T^{\mathcal{B}}$ , with leaf set  $\cup_{x \in X} A_x$ , is obtained from  $T$  and  $\mathcal{B}$  by identifying each leaf  $x$  of  $T$  with the root  $x$  on  $B_x$ . (See Figure 1 for an

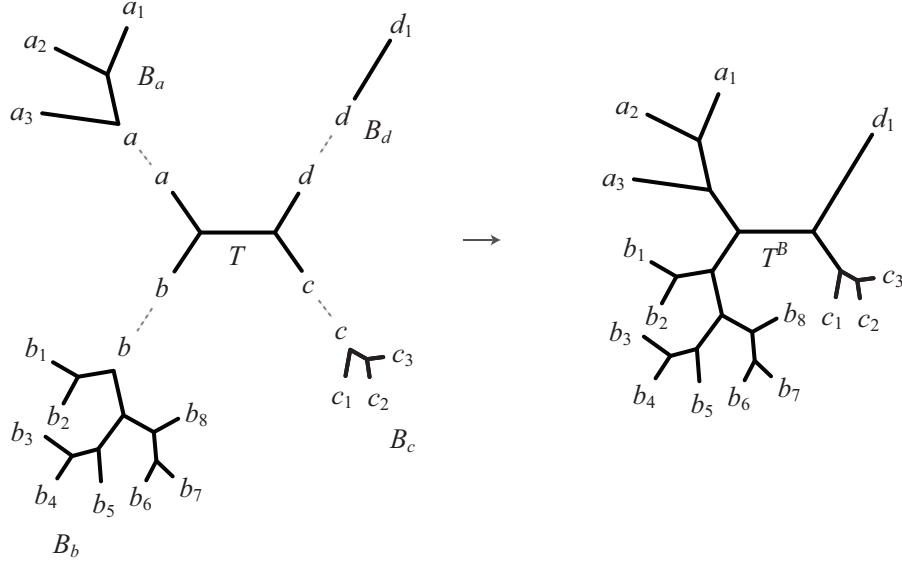


Figure 1: The fusion tree  $T^{\mathcal{B}}$  is constructed from a tree  $T$  with leaf set  $X = \{a, b, c, d\}$  and a set  $\mathcal{B} = \{B_a, B_b, B_c, B_d\}$  of fusion ends for  $X$ . The construction using  $\mathcal{B}$  could be applied to any of the quartet trees with leaf set  $X$ , yielding fusion trees differing by an NNI move from the  $T^{\mathcal{B}}$  shown here. This process underlies the extension of mimicking examples on small trees to larger ones.

example.)

If  $\mathcal{T}$  is a collection of trees each with leaf set  $X$ , and  $\mathcal{B} = \{B_x : x \in X\}$  is a set of fusion ends for  $X$ , let  $\mathcal{T}^{\mathcal{B}}$  be the multiset  $\mathcal{T}^{\mathcal{B}} = \{T^{\mathcal{B}} : T \in \mathcal{T}\}$  of fusion trees. The following propositions allow us to pass mimicking properties from small trees to large trees.

**Proposition 4.** *Suppose for a taxon set  $X$  that  $\mathcal{M}_{\mathcal{T}}^*$  weakly mimics  $\mathcal{M}_{\mathcal{S}}^*$ , and that  $\mathcal{B}$  is a set of fusion ends. Then  $\mathcal{M}_{\mathcal{T}^{\mathcal{B}}}^*$  weakly mimics  $\mathcal{M}_{\mathcal{S}^{\mathcal{B}}}^*$ .*

*Proof.* Any distribution  $q \in \mathcal{M}_{\mathcal{T}}^* \cap \mathcal{M}_{\mathcal{S}}^*$  arises from parameters on the trees in  $\mathcal{T}$ , as well as from parameters on the trees in  $\mathcal{S}$ . Extend these parameters to the trees in  $\mathcal{T}^{\mathcal{B}}$  and  $\mathcal{S}^{\mathcal{B}}$  by choosing a length and rate matrix for each edge of each tree in  $\mathcal{B}$ , and using these choices for the resulting edges in the individual fusion trees in  $\mathcal{T}^{\mathcal{B}}$  and  $\mathcal{S}^{\mathcal{B}}$ . Using the same mixing parameters as led to  $q$ , these parameters give rise to a distribution  $q^{\mathcal{B}} \in \mathcal{M}_{\mathcal{T}^{\mathcal{B}}}^* \cap \mathcal{M}_{\mathcal{S}^{\mathcal{B}}}^*$ .  $\square$

**Proposition 5.** *Let  $\mathcal{S} = \{T\}$ , be a single tree, and suppose that  $\mathcal{M}_{\mathcal{T}}^*$  strongly mimics (or*

completely mimics)  $\mathcal{M}_{\mathcal{S}}^*$ . Then for any set  $\mathcal{B}$  of fusion ends for  $X$ ,  $\mathcal{M}_{\mathcal{T}^{\mathcal{B}}}^*$  strongly mimics (or completely mimics)  $\mathcal{M}_{\mathcal{S}^{\mathcal{B}}}^*$ .

*Proof.* This follows from the same argument as for Proposition 4, with the additional observation that the parameters assigned to edges in the fusion ends can be varied arbitrarily. Since  $\mathcal{S}$  consists of a single tree, this will give a full dimensional set of distributions in  $\mathcal{M}_{\mathcal{S}^{\mathcal{B}}}^*$  which are mimicked by distributions in  $\mathcal{M}_{\mathcal{T}^{\mathcal{B}}}^*$ . If  $\mathcal{M}_{\mathcal{T}}^*$  completely mimics  $\mathcal{M}_{\mathcal{S}}^*$ , note that every distribution in  $\mathcal{M}_{\mathcal{S}^{\mathcal{B}}}^*$  arises from our construction so that  $\mathcal{M}_{\mathcal{T}^{\mathcal{B}}}^*$  completely mimics  $\mathcal{M}_{\mathcal{S}^{\mathcal{B}}}^*$ .  $\square$

These propositions allow the construction of explicit examples of mimicking behavior on large trees from those found on small trees. A typical result of this type, using quartet trees as a basis, is:

**Theorem 6.** Let  $\mathcal{T} = \{T_{12|34}, \dots, T_{12|34}\}$ , with  $r$  copies of the quartet tree  $T_{12|34}$ , and  $\mathcal{S} = \{T_{13|24}, \dots, T_{13|24}\}$ , with  $s$  copies of the quartet tree  $T_{13|24}$ . Let  $T$  and  $T'$  be trees with  $n \geq 4$  leaves that differ by an NNI move,  $\mathcal{T}' = \{T, \dots, T\}$ , with  $r$  copies of  $T$ , and  $\mathcal{S}' = \{T', \dots, T'\}$ , with  $s$  copies of  $T'$ .

If  $\mathcal{M}_{\mathcal{T}}^*$  weakly mimics  $\mathcal{M}_{\mathcal{S}}^*$ , then  $\mathcal{M}_{\mathcal{T}'}^*$  weakly mimics  $\mathcal{M}_{\mathcal{S}'}^*$ . Furthermore, if  $\mathcal{M}_{\mathcal{T}}^*$  strongly (or completely) mimics  $\mathcal{M}_{\mathcal{S}}^*$  and  $s = 1$ , then  $\mathcal{M}_{\mathcal{T}'}^*$  strongly (or completely) mimics  $\mathcal{M}_{\mathcal{S}'}^*$ .

*Proof.* Two trees differ by an NNI move if and only if they are obtained from applying fusions to two differing quartet trees. Hence, we can apply Propositions 4 and 5.  $\square$

In particular, Theorem 6 implies that if quartets give mimicking behavior, then we will have mimicking behavior on trees of arbitrary size. (Note conversely that Theorem 31 of Matsen et al. (2008) shows that the only way 2-class single-tree CFN mixtures can mimic CFN non-mixtures on large trees is through such a process applied to quartet over-parameterization.)

Consider now the general continuous-time model on a 4-leaf tree. With 5 edges, a distribution is specified by  $\approx 5\kappa^2$  numerical parameters. Since there are no linear tests for this model, and the probability distribution lies in a space of dimension  $\kappa^4 - 1$ , we expect that a mixture of more than  $\approx \kappa^2/5$  components will include an open subset of the probability simplex. Hence such a model is likely to display mimicking behavior. Thus some sort of mimicking seems unavoidable for even moderately sized mixtures. To illustrate, with DNA sequences and  $\kappa = 4$ , an unmixed model is specified by 63 parameters, so the 4-class mixture model has enough parameters that it is likely to include a full-dimensional subset and produce mimicking.

Note that mimicking of the sort produced by local over-parameterization as above need not be limited to that arising from quartet trees. With enough mixture components, for some models it may be possible for a mixture on a relatively small tree to mimic a distribution from another tree, differing by more than a single NNI move. This mimicking would again extend to larger trees, using the fusion process of Propositions 4 and 5.

From a practical perspective, however, mimicking through local over-parameterization seems unlikely to be much of an issue in most data analyses, since the mixture parameters leading to it require that the mixed processes differ only on a small part of the tree, and are identical elsewhere. Researchers studying biological situations in which this might be plausible should, however, be aware of the possibility.

Finally, we emphasize that we have not shown that local over-parameterization is the only possible source of mimicking. It would be quite interesting to have examples of mimicking of other sorts, or extensions of Theorem 31 of Matsen et al. (2008) to other models and more mixture components.

An early motivation for the study of linear invariants for phylogenetic models was that they are also invariants for mixture models on a single tree, and thus offered hope for determining tree topologies even under heterogeneous processes across sites. While poor practical performance (Huelsenbeck, 1995) even in the unmixed case led to their abandonment as an inference tool, they remain useful for theoretical purposes. However, among the commonly-studied phylogenetic models, the Jukes-Cantor (JC) and Kimura 2-parameter (K2P) models are the only ones which possess phylogenetically-informative linear invariants.

Štefankovič and Vigoda (2007a) used these linear invariants and the fact that they can be used to give linear tests, to show that if  $\mathcal{S}$  and  $\mathcal{T}$  are multisets each consisting of a single repeated  $n$ -leaf binary (fully-resolved) tree, and these trees are different, then  $\mathcal{M}_{\mathcal{T}}^+ \cap \mathcal{M}_{\mathcal{S}}^+ = \emptyset$ , regardless of the number of mixture components. We explore the extent to which these results can be extended to nonidentical tree mixtures for the JC and K2P models.

**Theorem 7.** *Consider the Jukes-Cantor and Kimura 2-parameter models. Let  $\mathcal{S} = \{T_1, \dots, T_1\}$  be a multiset of identical trees on  $X$ , and  $\mathcal{T}$  an arbitrary multiset of trees on  $X$ . If  $\mathcal{M}_{\mathcal{S}} \cap \mathcal{M}_{\mathcal{T}}^{++} \neq \emptyset$ , then for every four element subset  $K \subseteq X$ , and for all  $T \in \mathcal{T}$ , either  $T|_K$  is an unresolved (star) tree, or  $T|_K = T_1|_K$ . Thus all trees in  $\mathcal{T}$  have  $T_1$  as a binary resolution.*

*Furthermore, if all trees  $T \in \mathcal{T}$  are binary, and  $T_1 \notin \mathcal{T}$  then  $\mathcal{M}_{\mathcal{S}} \cap \mathcal{M}_{\mathcal{T}}^+ = \emptyset$ .*

Informally, the last statement of this theorem states that arbitrary multitree phylogenetic mixtures on fully-resolved trees cannot mimic mixtures on a single tree, unless that tree appears in some component of the mixture. Thus if one erroneously assumed such a mimicking distribution was from a single tree mixture, the single tree one would recover would in fact reflect the truth for at least one mixture component.

In the case that  $\mathcal{S} = \{T_1\}$ , so  $\mathcal{M}_{\mathcal{S}}$  is not a mixture but rather a standard model, for the JC and K2P models this again rules out any mimicking examples of the sort Matsen and Steel (2007) and Štefankovič and Vigoda (2007b) give for CFN, unless one allows zero length

branches. This clearly indicates the special nature that any such exceptional cases must have.

Our final theorem provides a construction of the special case of mimicking allowed by Theorem 7, for the Jukes-Cantor model, where  $\mathcal{T}$  contains nonbinary trees that are degenerations of the tree  $T$ .

**Theorem 8.** *Let  $T$  be a tree with internal vertex  $v$  which is adjacent to three other vertices  $u_1, u_2, u_3$ . For  $i = 1, 2$ , let  $T_i$  be the tree obtained from  $T$  by contracting the edge  $u_i v$ . Let  $\mathcal{S} = \{T\}$  and  $\mathcal{T} = \{T_1, T_2\}$ . Then, under the Jukes-Cantor model  $\mathcal{M}_{\mathcal{T}}^*$  completely mimics  $\mathcal{M}_{\mathcal{S}}^*$ .*

#### ACKNOWLEDGEMENTS

The work of Elizabeth Allman and John Rhodes is supported by the U.S. National Science Foundation (DMS 0714830), and that of Seth Sullivant by the David and Lucille Packard Foundation and the U.S. National Science Foundation (DMS 0954865).

This work was begun at the Institut Mittag-Leffler, during its Spring 2011 program ‘Algebraic Geometry with a View Towards Applications.’ The authors thank the Institute and program organizers for both support and hospitality.

#### REFERENCES

- Allman, E. S., S. Petrović, J. A. Rhodes, and S. Sullivant. 2011. Identifiability of two-tree mixtures for group-based models. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 8:710–722.
- Allman, E. S. and J. A. Rhodes. 2006. The identifiability of tree topology for phylogenetic models, including covarion and mixture models. *J. Comput. Biol.* 13:1101–1113.



- Cavender, J. A. and J. Felsenstein. 1987. Invariants of phylogenies in a simple case with discrete states. *J. of Class.* 4:57–71.
- Chai, J. and E. A. Housworth. 2011. On Rogers’s Proof of Identifiability for the GTR + Gamma + I Model. *Syst. Biol.* 60:713–718.
- Eriksson, N. 2005. Tree construction using singular value decomposition. Pages 347–358 *in* Algebraic Statistics for Computational Biology. Cambridge Univ. Press, New York.
- Evans, J. and J. Sullivan. 2012. Generalized mixture models for molecular phylogenetic estimation. *Syst. Biol.* 61:12–21.
- Evans, S. N. and T. P. Speed. 1993. Invariants of some probability models used in phylogenetic inference. *Ann. Statist.* 21:355–377.
- Galtier, N. and M. Gouy. 1998. Inferring pattern and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.* 15:871–879.
- Hendy, M. D. 1989. The relationship between simple evolutionary tree models and observable sequence data. *Systematic Zoology* 38:310–321.
- Huelsenbeck, J. P. 1995. Performance of phylogenetic methods in simulation. *Syst. Biol.* 44:17–48.
- Kolaczowski, B. and J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 431:980–984.
- Le, S., N. Lartillot, and O. Gascuel. 2008. Phylogenetic mixture models for proteins. *Phil. Trans. R. Soc. B.* 363:3965–3976.
- Matsen, F. A., E. Mossel, and M. Steel. 2008. Mixed-up trees: the structure of phylogenetic mixtures. *Bull. Math. Biol.* 70:1115–1139.

- Matsen, F. A. and M. A. Steel. 2007. Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Syst. Biol.* 56:767–775.
- Mossel, E. and E. Vigoda. 2005. Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science* 309:2207–2209.
- Mossel, E. and E. Vigoda. 2006. Limitations of Markov chain Monte Carlo algorithms for Bayesian inference of phylogeny. *Ann. Appl. Probab.* 16:2215–2234.
- Pagel, M. and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.* 53:571–581.
- Pagel, M. and A. Meade. 2005. Mixture models in phylogenetic inference. Pages 121–142 *in* *Mathematics of Evolution and Phylogeny* (O. Gascuel, ed.). Oxford University Press, Oxford.
- Rhodes, J. A. and S. Sullivant. 2012. Identifiability of large phylogenetic mixture models. *Bull. Math. Biol.* 74:212–231.
- Ronquist, F. R. and J. P. Huelsenbeck. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1574–1575.
- Steel, M. 2009. A basic limitation on inferring phylogenies by pairwise sequence comparisons. *J. Theoret. Biol.* 256:467–472.
- Steel, M. 2011. Can we avoid “SIN” in the house of “No Common Mechanism”? *Syst. Biol.* 60:96–109.
- Steel, M., L. Székely, and M. Hendy. 1994. Reconstructing trees when sequence sites evolve at variable rates. *J. Comput. Biol.* 1:153–163.
- Tuffley, C. and M. Steel. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bull. Math. Biol.* 59:581–607.

- Štefankovič, D. and E. Vigoda. 2007a. Phylogeny of mixture models: Robustness of maximum likelihood and non-identifiable distributions. *J. Comput. Biol.* 14:156–189.
- Štefankovič, D. and E. Vigoda. 2007b. Pitfalls of heterogeneous processes for phylogenetic reconstruction. *Syst. Biol.* 56:113–124.
- Wang, H. C., K. Li, S. E., and A. J. Roger. 2008. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC Evol Biol.* 8:331.
- Yang, Z. and D. Roberts. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. *Mol. Biol. Evol.* 12:451–458.
- Yap, V. and T. Speed. 2005. Rooting a phylogenetic tree with nonreversible substitution models. *BMC Evol. Biol.* 5:1–8.

## APPENDIX: MATHEMATICAL ARGUMENTS

To prove Theorem 2 we first handle the special case of 4-leaf trees. We need the following definition.

**Definition 9.** If  $P$  is a probability distribution for a  $\kappa$ -state phylogenetic model on a  $n$ -taxon tree, we view it as an  $n$ -dimensional  $\kappa \times \kappa \times \cdots \times \kappa$  tensor, or array, of probabilities,  $P = (p_{i_1 i_2 \dots i_n})$ , where the index  $i_l$  refers to the state at leaf  $l$ . Then given any bipartition of the leaves into non-empty subsets  $\{1, 2, \dots, n\} = A \sqcup B$ , the  $A|B$  *flattening* of  $P$  is the  $\kappa^{|A|} \times \kappa^{|B|}$  matrix  $\text{Flat}_{A|B}$  with the same entries as  $P$  but with rows indexed by state assignments to leaves in  $A$ , and columns indexed by state assignments to leaves in  $B$ .

**Lemma 10.** *Consider 4-leaf trees  $T_1$  with split 12|34, and  $T_2$  either the tree with split 13|24 or the star tree. Then the statement of Theorem 2 holds. That is,  $\mathcal{M}_{\mathcal{T}} \cap \mathcal{M}_{\mathcal{S}}^+ = \emptyset$  unless  $T_2$  is the star tree.*

*Proof.* Let  $p$  denote a probability distribution  $p \in \mathcal{M}_{\mathcal{T}}$ , which we consider as a 4-dimensional tensor. Consider the  $\{1, 2\}|\{3, 4\}$  flattening  $\text{Flat}_{12|34}(p)$ , which is a  $\kappa^2 \times \kappa^2$  matrix. From Allman and Rhodes (2006) or Eriksson (2005) it is known that if  $p \in \mathcal{M}_{\mathcal{T}}$  then the rank of  $\text{Flat}_{12|34}(p)$  is at most  $\kappa(\kappa - 1)$ .

On the other hand, if  $T_2 = 13|24$  and  $q \in \mathcal{M}_{\mathcal{S}}^+ = \mathcal{M}_{T_2}^+$ , then the matrix  $\text{Flat}_{12|34}(q)$  has a factorization as

$$\text{Flat}_{12|34}(q) = (M_1 \otimes M_2) \text{diag}(N) (M_3 \otimes M_4) \quad (1)$$

where  $M_i, 1 \leq i \leq 4$  are the transition matrices associated with the leaf edges in the tree, and  $N = \text{diag}(\boldsymbol{\pi})M_5$  where  $M_5$  is the transition matrix associated to the internal edge and we have assumed the tree root is at one end of that edge. Here  $\text{diag}(N)$  denotes a  $\kappa^2 \times \kappa^2$  diagonal matrix constructed with the entries of  $N$  on its diagonal in an appropriate order. By Lemma 1, all transitions matrices for the model  $\mathcal{M}_{T_2}$  are nonsingular. Thus the  $\kappa^2 \times \kappa^2$  matrices  $M_1 \otimes M_2$  and  $M_3 \otimes M_4$  are nonsingular. Also by Lemma 1, for the open model  $\mathcal{M}_{T_2}^+$ , the matrix  $\text{diag}(N)$  is nonsingular since all the entries of  $\boldsymbol{\pi}$  and  $M_5$  are nonzero. Thus if  $q \in \mathcal{M}_{T_2}^+$ ,  $\text{Flat}_{12|34}(q)$  has rank  $\kappa^2$ .

If  $T_2$  is the star tree, then formula (1) still holds if one sets  $M_5 = I$ . In this case the matrix  $\text{diag}(N)$  is singular, and the rank of  $\text{Flat}_{12|34}(q)$  is  $\kappa$ .

These conditions on the rank of  $\text{Flat}_{12|34}(q)$  now imply the desired conclusion.  $\square$

*Proof of Theorem 2.* If  $T_1$  is a refinement of  $T_2$ , then one checks that  $\mathcal{M}_{\mathcal{S}}^+ \subset \mathcal{M}_{\mathcal{T}}$ , by choosing the mixing weights as a standard unit vector, and setting edge lengths equal to zero on the edges appearing in  $T_1$  but not  $T_2$ .

So assume that  $T_1$  is not a refinement of  $T_2$ , yet  $\mathcal{M}_{\mathcal{T}} \cap \mathcal{M}_{\mathcal{S}}^+$  is non-empty. We may also assume that  $T_1$  is a binary tree, by passing to a refinement, as this only enlarges the mixture model. There exists a subset  $K$  of four taxa such that the induced quartet trees  $T_1|_K$  and  $T_2|_K$  are different. Marginalizing to  $K$ , since  $(\mathcal{M}_{\mathcal{T}})|_K = \mathcal{M}_{(\mathcal{T}|_K)}$  and  $(\mathcal{M}_{\mathcal{S}})|_K = \mathcal{M}_{(\mathcal{S}|_K)}$ , we have that  $\mathcal{M}_{(\mathcal{T}|_K)} \cap \mathcal{M}_{(\mathcal{S}|_K)}^+$  is non-empty.

Now, by Lemma 1 the transition matrices that arise in the resulting quartet trees will

be products of nonsingular matrices that either are the identity, or have all positive entries. Thus each quartet tree transition matrix is nonsingular and can have zero entries if, and only if, it is the product of identity matrices. We now apply Lemma 10 to deduce that all the edge lengths along the internal edge of  $T_2|_K$  must be zero. But this contradicts the fact that we were working with the open model  $\mathcal{M}_S^+$ .  $\square$

To prove Theorem 7, we recall a number of results about the JC and K2P models, including their descriptions in Fourier coordinates, and properties of linear invariants/tests for these models.

The JC, K2P (and K3P) models are group-based models, with a special structure governed by the finite abelian group  $G = \mathbb{Z}_2 \times \mathbb{Z}_2$ . We associate nucleotides with elements of this group via

$$A = (0, 0), \quad C = (0, 1), \quad G = (1, 0), \quad T = (1, 1).$$

The discrete Fourier transform (also called Hadamard conjugation in this context) (Hendy, 1989; Evans and Speed, 1993) is an invertible linear transformation that simplifies the parameterization of a group-based model. In Fourier coordinates,  $q_{g_1 \dots g_n}$ , the parametrization is described as follows: To each of the tree  $T$ 's splits  $A|B$  we associate a collection of parameters  $a_g^{A|B}$  where  $g \in G$ . Then

$$q_{g_1 \dots g_n} = \begin{cases} \prod_{A|B} a_{\sum_{i \in A} g_i}^{A|B} & \text{if } \sum g_i = 0 \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

**Proposition 11.** *Suppose that a transition matrix has the form  $\exp(Qt)$  where  $Q$  is a rate matrix for a  $\mathbb{Z}_2 \times \mathbb{Z}_2$  group-based model,  $t > 0$ , and  $Q$  defines an irreducible Markov chain. Then the Fourier parameters satisfy the constraints:*

$$\begin{aligned} a_A^{A|B} &= 1, \\ a_C^{A|B} &\geq a_G^{A|B} a_T^{A|B}, \\ a_G^{A|B} &\geq a_C^{A|B} a_T^{A|B}, \\ a_T^{A|B} &\geq a_C^{A|B} a_G^{A|B}, \end{aligned}$$

with  $a_C^{A|B}, a_G^{A|B}, a_T^{A|B} \in (0, 1)$ . When  $t = 0$ , all parameters equal 1.

Additionally, under the K2P model,  $a_G^{A|B} = a_T^{A|B}$ , and under the JC model  $a_C^{A|B} = a_G^{A|B} = a_T^{A|B}$ .

*Proof.* Let  $Q$  be a rate matrix of K3P format and  $H$  the associated  $4 \times 4$  Hadamard matrix, that is, for some  $\alpha, \beta, \gamma \geq 0$ ,  $\delta = -\alpha - \beta - \gamma$ ,

$$Q = \begin{bmatrix} \delta & \alpha & \beta & \gamma \\ \alpha & \delta & \gamma & \beta \\ \beta & \gamma & \delta & \alpha \\ \gamma & \beta & \alpha & \delta \end{bmatrix} \quad \text{and} \quad H = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}.$$

The Fourier coordinates for this model consist of the eigenvalues of the matrix  $\exp(Qt)$ . The matrix  $H$  consists of the eigenvectors of the matrix  $Q$ , and hence of  $\exp(Qt)$ . We compute that  $H^{-1}QH$  is the diagonal matrix  $\text{diag}(0, -2(\alpha + \gamma), -2(\beta + \gamma), -2(\alpha + \beta))$ . From this we deduce that the Fourier coordinates for this model are then

$$a_A^{A|B} = 1, \quad a_C^{A|B} = \exp(-2t(\alpha + \gamma)), \quad a_G^{A|B} = \exp(-2t(\beta + \gamma)), \quad a_T^{A|B} = \exp(-2t(\alpha + \beta)).$$

Since  $Q$  gives an irreducible Markov chain, at most one of  $\alpha, \beta$ , and  $\gamma$  can be zero, which implies that all of  $a_C^{A|B}, a_G^{A|B}, a_T^{A|B} < 1$  when  $t > 0$ . Furthermore, we see that the claimed inequalities hold, *e.g.*,

$$a_C^{A|B} = \exp(-2t(\alpha + \gamma)) \geq \exp(-2t(\alpha + 2\beta + \gamma)) = a_G^{A|B} a_T^{A|B}.$$

Note also that the K2P model consists of all rate matrices where  $\alpha = \gamma$ , which implies that  $a_G^{A|B} = a_T^{A|B}$ , and the JC models consists of all rate matrices where  $\alpha = \beta = \gamma$ , which implies that  $a_C^{A|B} = a_G^{A|B} = a_T^{A|B}$ .  $\square$

**Proposition 12.** Let  $T_1 = T_{12|34}$ ,  $T_2 = T_{13|24}$ , and  $T_3 = T_{14|23}$ . Then under the JC and K2P models, the polynomial

$$l(q) = q_{GGGG} - q_{GGTT}$$

satisfies the following properties:

1.  $l(q) = 0$  for all  $q \in \mathcal{M}_{T_1}$ ,
2.  $l(q) \geq 0$  for all  $q \in \mathcal{M}_{T_i}$ ,  $i = 2, 3$ ,
3.  $l(q) > 0$  for all  $q \in \mathcal{M}_{T_i}^+$ ,  $i = 2, 3$ , and
4. if  $q \in \mathcal{M}_{T_i}$ , for  $i = 2$  or  $3$ , and  $l(q) = 0$ , then the branch length of the internal edge is zero.

*Proof.* To evaluate the polynomial  $l(q)$ , we substitute for  $q$  the parametric expressions given in equation (2). Denoting parameters for trivial splits  $\{i\} | (\{1, 2, 3, 4\} \setminus \{i\})$  by  $a_g^i$ , for  $q \in \mathcal{M}_{T_1}$  we have

$$l(q) = q_{GGGG} - q_{GGTT} = a_G^1 a_G^2 a_G^3 a_G^4 a_A^{12|34} - a_G^1 a_G^2 a_T^3 a_T^4 a_A^{12|34}.$$

Since  $a_G^{A|B} = a_T^{A|B}$  in the JC and K2P models, the first claim follows.

If  $q \in \mathcal{M}_{T_2}$ , to establish the remaining claims note

$$l(q) = q_{GGGG} - q_{GGTT} = a_G^1 a_G^2 a_G^3 a_G^4 a_A^{13|24} - a_G^1 a_G^2 a_T^3 a_T^4 a_C^{13|24}.$$

Since  $a_G^{A|B} = a_T^{A|B}$  for the JC and K2P models, and  $a_A^{13|24} = 1$ , this expression factors as

$$l(q) = a_G^1 a_G^2 a_G^3 a_G^4 (1 - a_C^{13|24}).$$

By Proposition 11 all  $a_g^{A|B} \in (0, 1]$ , so  $l(q) \geq 0$ . Moreover, if all branch lengths are strictly positive, so is  $l(q)$ . On the other hand, the only way this expression can equal zero with  $q \in \mathcal{M}_{T_2}$  is if  $a_C^{13|24} = 1$ . But then Proposition 11 implies the length of the internal branch is zero.

Similar arguments show the claims for  $T_3$ . □

*Proof of Theorem 7.* Let  $K$  be any four element subset of the taxa. If  $\mathcal{M}_{\mathcal{T}} \cap \mathcal{M}_{\mathcal{S}}^+ \neq \emptyset$ , then when we marginalize to mixture models on the leaf set  $K$  the corresponding intersection is also non-empty. Since the claims of the theorem concern quartets, it suffices to restrict attention to the case of  $n = 4$  taxa.

First suppose that the tree  $T_1$  is fully-resolved. By symmetry we may assume it is  $T_{12|34}$ . By Proposition 12,  $l(q) = 0$  if  $q \in \mathcal{M}_{T_{12|34}}$ , while  $l(q) > 0$  if  $q \in \mathcal{M}_{T_{13|24}}^+$  or  $\mathcal{M}_{T_{14|23}}^+$ . By the linearity of  $l$ , this implies  $l(q) = 0$  if  $q \in \mathcal{M}_{\mathcal{S}}$ , while  $l(q) > 0$  for  $q \in \mathcal{M}_{\mathcal{T}}^{++}$  provided  $\mathcal{T}$  contains at least one of the resolved trees  $T_{13|24}$  or  $T_{14|23}$ . This implies that if  $q \in \mathcal{M}_{\mathcal{S}} \cap \mathcal{M}_{\mathcal{T}}^{++}$ , then no quartet incompatible with tree  $T_1$  can appear among the trees of  $\mathcal{T}$ .

If  $T_1$  is the star tree, then from each of its three resolutions we obtain inequalities analogous to those for  $l(q)$ . These imply that  $\mathcal{T}$  can only contain star trees.

Finally, in the case that all  $T \in \mathcal{T}$  are binary and  $T_1 \notin \mathcal{T}$ , if  $q \in \mathcal{M}_{\mathcal{S}} \cap \mathcal{M}_{\mathcal{T}}^+$  then by replacing  $\mathcal{T}$  by a subset  $\mathcal{T}'$  we have  $q \in \mathcal{M}_{\mathcal{S}} \cap \mathcal{M}_{\mathcal{T}'}^{++}$ . From the argument above it follows that for all  $T \in \mathcal{T}'$  and all quartets  $K$ ,  $T|_K = T_1|_K$ . Thus we obtain the contradiction that  $T = T_1$ , and conclude no such  $q$  exists.  $\square$

*Proof of Theorem 8.* We first consider the case that  $T$  is a 3-leaf tripod tree, and  $T_1$  and  $T_2$  are two of its contractions where one leaf has become an internal vertex.

The model on a 3-leaf tree under the JC model has precisely 3 nontrivial Fourier parameters, one per edge. We set the parametrization of that model, with edge parameters  $a, b, c \in (0, 1]$ , equal to the one for the mixture on  $T_1$  and  $T_2$ , with edge parameters  $d, e$  and  $f, g$  respectively, and mixing parameter  $\pi$ . This gives us, for fixed  $a, b, c$ , the following system of 4 equations in 5 unknowns:

$$\begin{aligned} ab &= (1 - \pi)d + \pi f, \\ ac &= (1 - \pi)de + \pi g, \\ bc &= (1 - \pi)e + \pi fg, \\ abc &= (1 - \pi)de + \pi fg. \end{aligned}$$

It is not difficult to see that the values

$$d = 0, \quad e = bc, \quad f = b, \quad g = c, \quad \pi = a \tag{3}$$



give a solution to this system of equations. For the open models, however, we seek solutions where  $d, e, f, g, \pi \in (0, 1)$  for fixed  $a, b, c \in (0, 1)$ . A computation of the Jacobian of the system of equations at the values in equations (3) allows us to apply the implicit function theorem, and treat  $d$  as an independent variable in a neighborhood of the above solution. Hence, if we perturb  $d$  to  $d'$ , with  $0 < d' \ll 1$ , we obtain parameters in  $(0, 1)$  solving the system of equations. This shows that there is complete mimicking for the open models in the 3-leaf case.

Finally we apply Proposition 5: Since any trees of the type specified in the statement of the theorem can be obtained by attaching fusion ends to the tripod tree and its two degenerations, we deduce the general result.  $\square$